SELF-RECONSTRUCTION DYNAMICS FOR AUTOENCODER RECONSTRUCTION REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard autoencoder (AE) inference uses a single encoder–decoder pass, although the resulting latent representation need not be optimal for each sample under the fixed decoder. This raises a natural question: *Can a trained AE itself reveal information useful for improving its own reconstruction?* We investigate this question by repeatedly applying a frozen AE to its own reconstruction, producing transient image- and latent-space trajectories that we call *Self-Reconstruction Dynamics (SRD)*. Although repeated self-reconstruction degrades fidelity in the AEs studied here, the resulting SRD contains useful sample-specific information for correcting the initial reconstruction. We therefore propose *SRD-guided Reconstruction Refinement (SRD-RR)*, which predicts a latent correction from a short SRD while keeping the AE frozen and requiring no per-sample test-time optimization. We also introduce an *MSE recovery ratio (MSE-recov)* relative to an empirical decoder-optimized reference. Across six image datasets, SRD-RR recovers an average of 38.6% of the empirically recoverable MSE gap with one trajectory transition and 45.3% with two. A two-transition variant trained without direct access to the original images, using an SRD-derived pseudo-target, achieves 40.7% recovery and a 1.74 dB average PSNR gain. Removing trajectory information substantially reduces the gain, while cross-sample trajectory assignment causes severe degradation, showing that the useful SRD information is strongly sample-specific. Nonlinear SRD-conditioned refinement also consistently outperforms both fixed and trained linear latent correction. We further evaluate SRD on a pretrained DINOv2-based representation autoencoder (RAE) with substantially different latent dynamics. SRD conditioning again improves a matched trajectory-free predictor, showing that trajectory information remains useful in this different AE setting. However, pixel-MSE latent refinement exposes a strong mismatch between pixel fidelity and perceptual quality, while the SRD-derived pseudo-target substantially mitigates this perceptual degradation. Together, these results establish SRD as a useful sample-specific signal for reconstruction refinement, while showing that the choice of refinement objective determines how this information translates into pixel and perceptual quality.

1 INTRODUCTION

Autoencoders (AEs) map high-dimensional observations to latent representations and reconstruct them through a decoder (Hinton & Salakhutdinov, 2006). Reconstruction quality remains important in modern latent generative models, including VAEs and latent diffusion models (Kingma & Welling, 2013; Rombach et al., 2022), while recent representation autoencoders (RAEs) emphasize both reconstruction fidelity and representation quality (Zheng et al., 2026). Standard AE inference is essentially one-shot: an encoder predicts a latent representation that is immediately decoded. Yet this representation need not yield the best reconstruction of each sample under the fixed decoder. Latent inversion and learned inference refinement likewise show that sample-specific refinement can improve an initial encoder prediction (Zhu et al., 2016; Xia et al., 2022; Cremer et al., 2018; Kim et al., 2018). This motivates a basic question: *Can a trained AE itself reveal information useful for further improving reconstruction?*

We investigate this question by repeatedly applying a frozen AE to its own reconstruction. This produces transient image- and latent-space trajectories, which we call *Self-Reconstruction Dynam-*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ics (SRD). Unlike optimization trajectories designed to approach the original image, repeated self-reconstruction in the AEs studied here progressively degrades fidelity. Nevertheless, the resulting trajectory contains useful sample-specific information for correcting the initial reconstruction. Rather than using later iterates as improved outputs, we therefore treat the trajectory itself as a sample-specific signal produced by the frozen AE.

Based on this observation, we propose *SRD-guided Reconstruction Refinement (SRD-RR)*, which uses a short SRD to predict a decoder-compatible latent correction while keeping the encoder and decoder frozen and requiring no per-sample test-time optimization. Trajectory removal and cross-sample shuffling show that the gain depends on the correspondence between each sample and its own SRD, rather than only on the initial latent representation or generic AE drift. We also introduce an x-free variant trained with an SRD-derived pseudo-target; once the self-reconstruction quantities are generated, refinement training and inference require no direct access to the original image x. To quantify improvement relative to the remaining capability of the fixed decoder, we introduce an *MSE recovery ratio (MSE-recov)* based on an empirical decoder-optimized reference obtained by per-sample latent optimization. It measures the fraction of the empirically recoverable MSE gap closed by a refinement method.

Across six image datasets, nonlinear SRD-conditioned refinement consistently improves reconstruction and outperforms both fixed and trained linear latent extrapolation. Most useful information is captured by the first few trajectory transitions, and the improvement remains consistent across random seeds and AE architectures. An additional experiment with a pretrained DINOv2-based RAE (Zheng et al., 2026) shows that SRD remains informative despite substantially different latent dynamics, while the refinement objective strongly affects the balance between pixel fidelity and perceptual preservation.

Contributions. Our contributions are: (1) we identify SRD as a sample-specific reconstruction signal arising naturally from repeated application of a frozen AE, even though repeated self-reconstruction degrades fidelity; (2) we propose SRD-RR, including an x-free variant, to exploit a short SRD for decoder-compatible latent correction without per-sample test-time optimization; and (3) we introduce MSE-recov for decoder-relative evaluation and systematically analyze the informativeness, sample specificity, geometry, and robustness of SRD across AEs.

2 Related Work

Iterated autoencoding and AE dynamics. Iterative or repeated application of autoencoder-based mappings has been studied in several contexts, including generative sampling (Bengio et al., 2013; Creswell et al., 2016), denoising (Bengio et al., 2013), latent dynamics (Nagano et al., 2020), associative memory (Radhakrishnan et al., 2020), and restoration (Lee et al., 2025). Regularized AEs have also been shown to induce reconstruction vector fields related to local data structure (Alain & Bengio, 2014). More recent approaches such as Epsilon-VAE (Zhao et al., 2025) and Cold Diffusion (Bansal et al., 2023) use iterative processes explicitly designed for reconstruction or inversion. SRD-RR differs in that it does not use the iteration itself as the improved output; instead, it exploits the transient trajectory naturally produced by a frozen AE as conditioning information for a separate latent correction.

Latent inversion and refinement. Latent inversion improves reconstruction under a fixed generator or decoder, commonly through per-image latent optimization (Zhu et al., 2016; Xia et al., 2022). Similar per-instance latent optimization has been studied in neural image compression (Campos et al., 2019; Yang et al., 2020), while iterative and semi-amortized inference refine initial encoder predictions in VAEs (Cremer et al., 2018; Marino et al., 2018; Kim et al., 2018). Learned iterative inversion methods such as ReStyle predict successive latent corrections (Alaluf et al., 2021), while gradient-free decoder inversion has been explored using fixed-point and inertial iterations (Hong et al., 2024).

Positioning relative to prior work. The key distinction of SRD-RR is the source of the refinement information. Existing inversion and refinement methods generate updates as part of a target-directed procedure. In SRD-RR, the trajectory is generated before any refinement by simply reapplying the frozen AE to its own reconstruction; it is not designed to approach the target and may instead move away from it. SRD-RR interprets this naturally occurring trajectory as sample-specific side information and predicts a separate latent correction without per-sample optimization at inference

time. To our knowledge, prior work has not used such a naturally generated self-reconstruction trajectory for this purpose.

3 SRD-GUIDED RECONSTRUCTION REFINEMENT (SRD-RR) FRAMEWORK

3.1 PROBLEM FORMULATION AND SELF-RECONSTRUCTION DYNAMICS

Let E and D denote the encoder and decoder, respectively, and define $\mathrm{AE} = D \circ E$. For an input image x, repeated application of the same AE gives

$$y_1 = \mathrm{AE}(x), \qquad y_{k+1} = \mathrm{AE}(y_k), \quad k \geq 1,$$

with latent representations

$$z_0 = E(x), \qquad z_k = E(y_k), \quad k \geq 1.$$

Thus, $y_1 = D(z_0)$ and $y_{k+1} = D(z_k)$. We define the reverse-step vectors

$$dy_k = y_k - y_{k+1}, \qquad dz_k = z_k - z_{k+1},$$

which point opposite to the direction of repeated autoencoding.

We refer to the sequences $\{y_k\}_{k\geq 1}$ and $\{z_k\}_{k\geq 1}$ as the *Self-Reconstruction Dynamics (SRD)*. SRD-guided Reconstruction Refinement (SRD-RR) uses information from this trajectory to obtain a reconstruction $\hat{x}$ that improves upon the one-shot reconstruction y_1.

We consider post-hoc refinement from y_1. At refinement inference, only y_1 and the frozen AE are available; x and z_0 are not retained. Thus, SRD-RR uses the observable latent trajectory

$$z_1,\ dz_1,\ dz_2, \ldots,$$

generated solely by repeatedly applying the fixed AE to y_1.

3.2 EMPIRICAL DECODER-OPTIMIZED REFERENCE AND EVALUATION

To quantify the reconstruction headroom of the fixed decoder, we optimize a latent representation for each sample while keeping D fixed:

$$z^* = \arg\min_z \|D(z) - x\|_2^2, \tag{1}$$

initialized from $z_0 = E(x)$. This follows standard latent-optimization practice in image inversion (Zhu et al., 2016; Xia et al., 2022).

We denote $y^* = D(z^*)$ as the *empirical decoder-optimized reference*. Because equation 1 is solved numerically and does not guarantee a global optimum, this is an empirical reconstruction reference rather than a theoretical bound.

For a refined reconstruction $\hat{x}$, we define the *MSE recovery ratio* (MSE-recov) as

$$\text{MSE-recov}(\hat{x}) = \frac{\mathrm{MSE}(y_1, x) - \mathrm{MSE}(\hat{x}, x)}{\mathrm{MSE}(y_1, x) - \mathrm{MSE}(y^*, x)}, \tag{2}$$

A value of 0 corresponds to the baseline reconstruction y_1, 1 to the empirical decoder-optimized reference y^*, and a negative value indicates degradation from y_1. Values above 1 are possible for methods not constrained to the output range of the fixed decoder, as they may achieve lower MSE than the decoder-optimized reference $y^* = D(z^*)$. In our experiments, this occurs only for the image-space M3 on two datasets. MSE-recov is computed independently for each sample before averaging over the dataset.

For latent-space geometry analysis, we additionally define the *empirical correction vector*

$$v^* = z^* - z_1, \tag{3}$$

which provides a sample-specific reference direction for comparison with the SRD reverse-step vectors dz_k.

Table 1: Summary of evaluated models, supervision, and x-free status.

Model	Description	Definition	Target	x-free
Baseline and image-space SRD models				
M1	Baseline AE	y_1	–	–
M2	Repeated AE	y_K	–	–
M3	Trained Nonlinear Image Corrector	$\tilde{x} = y_1 + \Delta x, \quad \Delta x = M_3(y_1, dy_1, \ldots, dy_{K-1})$	x	–
M3-AE	Corrector + AE	$\mathrm{AE}(M3)$	x	–
M4	Linear Image Extrapolation	$\tilde{x}_{\mathrm{lin}} = y_1 + \alpha(y_1 - y_2)$	–	✓
M4-AE	Linear Image Extrapolation + AE	$\mathrm{AE}(M4)$	–	✓
Linear latent-space SRD models				
M5	Linear Latent Extrapolation	$D(\tilde{z}_{\mathrm{lin}}), \quad \tilde{z}_{\mathrm{lin}} = z_1 + \beta_1 dz_1$	–	✓
M5-AE	Linear Latent Extrapolation + AE	$\mathrm{AE}(M5)$	–	✓
M5-T	Trained Linear Trajectory Correction	$D(\tilde{z}^*_{\mathrm{lin}}), \quad \tilde{z}^*_{\mathrm{lin}} = z_1 + \sum_{k=1}^{K-1} \beta^*_k dz_k$	x	–
Nonlinear latent-space SRD models; $D(z_1 + \gamma\Delta z)$				
M6-LO	Latent-Only Corrector	$\Delta z = M_{6\mathrm{LO}}(z_1)$	x	–
M6	SRD-RR	$\Delta z = M_6(z_1, dz_1, \ldots, dz_{K-1})$	x	–
M6-Lobj	Latent-Objective SRD-RR	$\Delta z = M_{6\mathrm{LAT}}(z_1, dz_1, \ldots, dz_{K-1})$	z^*	–
M6-xFree	x-free SRD-RR	$\Delta z = M_{6\mathrm{DEP}}(z_1, dz_1, \ldots, dz_{K-1})$	$\tilde{x}_{\mathrm{lin}}$	✓
M6-Shuf	Shuffled-Trajectory SRD-RR ⋆	$\Delta z^{(i)} = M_6(z_1^{(i)}, dz_1^{(\pi(i))}, \ldots, dz_{K-1}^{(\pi(i))})$	x	–
M7	SRD-RR + $\tilde{z}_0$ †	$\Delta z = M_7(z_1, dz_1, \ldots, dz_{K-1}, \tilde{z}_0)$	x	–
M8	SRD-RR + $\tilde{z}_{0,\mathrm{lin}}$ ‡	$\Delta z = M_8(z_1, dz_1, \ldots, dz_{K-1}, \tilde{z}_{0,\mathrm{lin}})$	$\tilde{x}_{\mathrm{lin}}$	✓
Empirical Decoder-optimized Reference		$D(z^*)$	x	–

⋆: Test-time shuffling uses the trained M6 predictor without retraining.
†: $\tilde{z}_0 = E(\tilde{x})$, where $\tilde{x}$ is the output of M3.
‡: $\tilde{z}_{0,\mathrm{lin}} = E(\tilde{x}_{\mathrm{lin}})$, where $\tilde{x}_{\mathrm{lin}}$ is the output of M4 (x-free).

3.3 Model Variants and x-Free Refinement

Table 1 summarizes all evaluated models. SRD-RR predicts a latent correction from the observable trajectory as

$$\hat{z} = z_1 + \gamma\Delta z_\theta(z_1, dz_1, \ldots, dz_{N_{\mathrm{tr}}}), \qquad \hat{x} = D(\hat{z}).$$

Here, $N_{\mathrm{tr}} = K - 1$ denotes the number of observed SRD transitions, and γ is a learnable scalar. M6 minimizes $\|\hat{x} - x\|_2^2$, whereas M6-xFree replaces x with the SRD-derived pseudo-target $\tilde{x}_{\mathrm{lin}} = 2y_1 - y_2$, i.e., the M4 output with $\alpha = 1$. All latent-space correctors are anchored at $z_1 = E(y_1)$, so zero correction yields $D(z_1) = y_2$ rather than y_1. Unlike M3 and M4, which are anchored directly at y_1, they must first recover the $y_1 \to y_2$ degradation before improving upon the one-shot baseline.

We call a refinement method x-free if, once y_1 has been obtained, neither its training nor refinement inference requires direct access to x or any ground-truth-dependent quantity. Under this definition, M4, M4-AE, M5, M5-AE, M6-xFree, and M8 are x-free.

4 Experiments

4.1 Common Experimental Setup

Datasets. We evaluate SRD-RR on CIFAR-10 (Krizhevsky, 2009), SVHN (Netzer et al., 2011), FashionMNIST (Xiao et al., 2017), MNIST (LeCun et al., 1998), ImageNet100 (Deng et al., 2009; Russakovsky et al., 2015), and CelebA (Liu et al., 2015). We abbreviate FashionMNIST and ImageNet100 as FMNIST and ImageNet, respectively.

Dataset splits are summarized in Table 2. For the first four datasets, the original training set is split 9:1 for training and validation, with the official test set used for evaluation. For ImageNet, the original training set is split similarly, with 5,000 fixed samples from the original validation set used for testing. For CelebA, we split the original training set 9:1 and use 5,000 fixed samples from the official test set. For ImageNet and CelebA, the nonlinear trajectory predictors (M6 family, M7, and M8) use fixed subsets of 20,000 training and 2,000 validation samples because of the storage cost of multi-step latent trajectories. The empirical latent z^* is used only for evaluation and analysis, except in M6-Lobj, where z^* computed from training samples serves as the regression target. Image sizes and pre-processing are described in Appendix C.1.

Evaluation. We use MSE-recov and PSNR as the primary reconstruction metrics. Both are computed independently for each test sample and then averaged within each dataset. Cross-dataset

Table 2: Dataset splits used in the experiments.

Split	CIFAR-10	SVHN	FashionMNIST	MNIST	ImageNet100	CelebA
Train	45,000	65,932	54,000	54,000	114,021	146,493
Val	5,000	7,325	6,000	6,000	12,668	16,277
Test	10,000	26,032	10,000	10,000	5,000	5,000

results are macro-averaged over the six datasets. Inference latency is measured on 500 ImageNet images with batch size 1 on an NVIDIA RTX 6000 Pro GPU, including all computation from x to the final output but excluding model loading. Details of preprocessing, model architecture, training, and normalization are provided in Appendix C.

4.2 Experiment 1: Image-Space Correction

We first test whether image-space SRD contains reconstruction-relevant information. M1 and M2 denote one-shot and repeated AE reconstruction, respectively. M3 learns a nonlinear residual correction Δx from $(y_1, dy_1, \ldots, dy_{K-1})$, whereas M4 performs linear image-space extrapolation. M4 is related in form to unsharp masking (Polesel et al., 2000), although its residual is induced by self-reconstruction rather than an explicit blurring filter. We fix $\alpha = 1.0$, corresponding to direct linear extrapolation of the first self-reconstruction step. A validation-set sweep confirms that this choice performs well across all datasets (Appendix Fig. A1). M3-AE and M4-AE pass the corrected images through the AE once more to test whether the improvement survives re-encoding.

4.3 Experiment 2: Linear Latent-Space Correction

We evaluate fixed linear latent extrapolation M5 and trained linear trajectory correction M5-T (Table 1) to test whether latent SRD provides a useful linear correction direction. Validation sweeps consistently favor $\beta \approx 2.0$ across all six datasets; we therefore use $\beta = 2.0$ for final test evaluation. Notably, the independently trained M5-T also learns first-step coefficients close to 2.0. Because M5 is training-free, computationally efficient, decoder-compatible, and x-free, we additionally study repeated application of M5 in Appendix A.

4.4 Experiment 3: Geometry of SRD

We analyze the local coherence, curvature, and reconstruction relevance of latent SRD using $\cos(dz_k, dz_{k+1})$, $\cos(dz_0, dz_k)$, and $\cos(v^*, dz_k)$. These quantify successive directional consistency, deviation from the initial trajectory direction, and alignment with the empirical correction direction, respectively. The quantities dz_0 and v^* are used only for analysis.

4.5 Experiment 4: Nonlinear SRD-RR and Robustness

We compare nonlinear SRD-RR M6 with several variants and ablations. M6-LO serves as a trajectory-free baseline that uses only z_1, allowing us to isolate the contribution of the SRD trajectory. We also evaluate the latent-objective M6-Lobj, x-free M6-xFree, shuffled-trajectory M6-Shuf, and auxiliary-guide variants M7 and M8 (Table 1), using $N_{\text{tr}} \in \{1, 2, 4\}$. For M6-Shuf, we keep the trained predictor and each sample's z_1 fixed and replace its trajectory with that of another sample at evaluation time.

We additionally evaluate robustness to training randomness and AE architecture. Experiments on CIFAR-10 and ImageNet are repeated with five random seeds, and three AE configurations with different depths and channel capacities are evaluated.

4.6 Experiment 5: Extension to a Pretrained Representation Autoencoder

To examine SRD in a substantially different autoencoder, we use a pretrained DINOv2-based representation autoencoder (RAE) to address two questions: (1) whether SRD remains informative under qualitatively different latent dynamics, and (2) how the refinement objective affects pixel and perceptual reconstruction quality. Both RAE components are kept frozen in our experiments. The SRD-RR predictor uses the same predictor backbone as in the ImageNet experiments, with RAE-specific settings described in Appendix D. We use 20,000/2,000/5,000 images for predictor training, validation, and testing, respectively, and evaluate M1, repeated reconstruction (M2), M4, M5, M6-LO, M6 ($N_{\text{tr}} = 1, 2$), and M6-xFree ($N_{\text{tr}} = 1, 2$). Here, M6-LO serves as the matched trajectory-free baseline for isolating the contribution of SRD conditioning.

Table 3: MSE recovery ratio (%) of image-space SRD methods.

Model	Description	CIFAR-10	SVHN	FMNIST	MNIST	ImageNet	CelebA
M1	y_1			0			
M2	y_2	-143.02	-197.19	-157.09	-158.24	-173.95	-172.62
	y_3	-346.00	-488.55	-371.01	-360.08	-429.96	-412.22
	y_5	-917.10	-1324.69	-961.63	-849.91	-1184.02	-1086.96
M3	$\tilde{x}$	116.87	103.86	87.09	96.20	51.54	98.73
M3-AE	AE(M3)	-8.93	-6.16	-5.38	-5.37	10.54	-14.03
M4 ($\alpha = 1.0$)	$\tilde{x}_{\text{lin}}$	52.72	68.33	61.88	63.75	63.36	63.09
M4-AE	AE(M4)	-14.52	-16.64	-21.05	-24.17	-14.96	-23.05
D-optimized Reference	$y^* = D(z^*)$			100			

Table 4: MSE recovery ratio (%) of linear latent-space correctors and estimated β_1^*.

Model	N_{tr}	CIFAR-10	SVHN	FMNIST	MNIST	ImageNet	CelebA
M5 ($\beta = 2.0$)	1	24.97	34.58	17.59	12.55	33.09	24.96
M5-AE	1	-40.84	-53.22	-54.55	-55.69	-38.46	-66.18
M5-T	1	24.99	33.89	18.03	13.66	32.15	19.42
	2	32.48	43.79	29.04	28.96	27.36	4.64
	3	32.52	43.36	29.68	31.51	24.36	31.10
	4	32.53	44.75	29.73	31.59	16.89	28.23
M5-T: estimated β_1^*	$(N_{tr} = 1)$	1.983	1.817	1.910	1.868	2.077	2.418

Evaluation uses MSE, PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and rFID-5k, computed as the Fréchet Inception Distance (FID) (Heusel et al., 2017) between the original and reconstructed image distributions on the fixed 5,000-image test set. Additional RAE-specific implementation and evaluation details are provided in Appendix D.

5 Results

5.1 Image-Space Correction

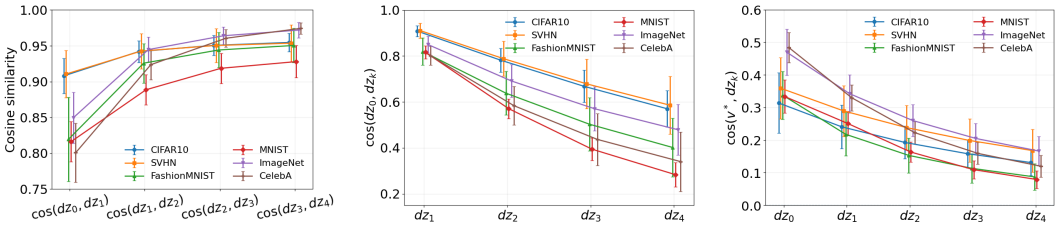
Table 3 summarizes the image-space results. Repeated autoencoding (M2) progressively degrades reconstruction as iteration increases, whereas both learned (M3) and linear (M4) correction substantially improve the baseline reconstruction. M3 achieves an average MSE-recov of 92.4% and exceeds the empirical decoder-optimized reference on two datasets. The auxiliary M4 sweep peaks near $\alpha = 1.0$ across datasets (Appendix Fig. A1), supporting our fixed choice $\alpha = 1.0$. However, these gains largely vanish after re-encoding: M3-AE and M4-AE perform at or below the baseline in most cases. Thus, strong image-space improvement does not necessarily correspond to a decoder-compatible latent representation.

5.2 Linear Latent-Space Correction

Table 4 shows the linear latent-space results. M5 improves reconstruction on all six datasets. The validation sweep consistently favors $\beta \approx 2.0$ (Appendix Fig. A1), consistent with the first coefficient independently learned by M5-T (Table 4). Additional transitions improve M5-T on the four smaller datasets, with diminishing gains, but produce dataset-dependent degradation or fluctuations on ImageNet and CelebA. M5-AE again falls below the baseline, indicating that the encoder does not preserve the refined latent after decoding and re-encoding. Repeated application of M5 provides no additional benefit beyond the first correction (Appendix Fig. A2), consistent with the changing direction of the latent SRD. These results motivate a nonlinear correction that explicitly exploits the short SRD.

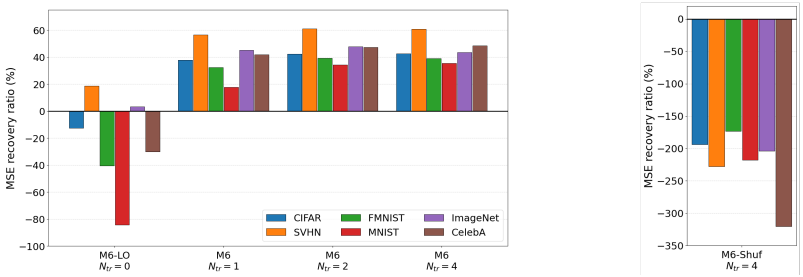
5.3 Geometry of SRD

Figure 1 characterizes the latent-space geometry of SRD. Successive reverse-step vectors have high cosine similarity, indicating strong local directional coherence, while their alignment with the initial direction decreases over time, revealing a curved rather than globally linear trajectory (Appendix Fig. A3). The observable dz_k also exhibit moderate positive alignment with the empirical cor-



(a) $\cos(dz_k, dz_{k+1})$ (b) $\cos(dz_0, dz_k)$ (c) $\cos(v^*, dz_k)$

Figure 1: Geometric properties of the latent Self-reconstruction Dynamics (SRD). Error bars indicate standard deviations within each dataset.



(a) Trajectory contribution (b) Trajectory shuffling

Figure 2: Effect of trajectory information and sample correspondence on nonlinear SRD-RR models.

rection vector v^*, particularly at early steps. Thus, early SRD contains information related to the reconstruction-improving direction without directly coinciding with it.

5.4 Nonlinear SRD-RR and Sample Specificity

Figure 2 shows that M6 consistently outperforms M6-LO, demonstrating that explicitly providing SRD features improves correction prediction over using z_1 alone. Most of the gain is captured by the first few steps and largely saturates by $N_{\mathrm{tr}} = 2$. Cross-sample trajectory shuffling causes severe degradation on every dataset. The benefit therefore depends on the correspondence between each sample and its own SRD, rather than on generic AE drift.

5.5 Overall Performance, Efficiency and Robustness

Table 5 summarizes the principal models in terms of MSE recovery. For M5–M8, zero correction yields $D(z_1) = y_2$, so they must recover the additional self-reconstruction degradation before surpassing y_1. At $N_{\mathrm{tr}} = 2$, M6 outperforms the trained linear corrector M5-T on all six datasets, supporting the benefit of nonlinear SRD-conditioned correction. The x-free M6-xFree improves reconstruction on all six datasets, achieving 40.7% average MSE recovery using only quantities derived from self-reconstruction. To complement this normalized evaluation, Appendix Table A1 reports absolute reconstruction performance in PSNR. M6-xFree improves PSNR by an average of $1.74\,\mathrm{dB}$ over the one-shot AE baseline, compared with $2.01\,\mathrm{dB}$ for the supervised M6 ($N_{\mathrm{tr}} = 2$). The baseline AE requires $0.257\,\mathrm{ms}$ per ImageNet image. M5 and M6-xFree ($N_{\mathrm{tr}} = 2$) require approximately $2.9\times$ and $6.3\times$ the baseline runtime, respectively, whereas computing the empirical decoder reference $D(z^*)$ requires approximately $338\,\mathrm{ms}$ ($\sim 1300\times$). This reference is not required during SRD-RR inference.

Reconstruction examples and error maps are shown in Appendix B. Across five random seeds and multiple AE architectures, SRD-RR consistently improves reconstruction across all settings. Although the exact gains vary, the overall trend is preserved across initializations and changes in AE depth and channel capacity (Figure A4 in Appendix).

5.6 Extension to a Pretrained Representation Autoencoder (RAE)

Table 6 summarizes the RAE results. The key comparison for SRD informativeness is M6 versus M6-LO, which uses the same predictor backbone and pixel-MSE objective but removes the trajectory input. Adding one SRD transition improves all five evaluated metrics, including PSNR by $0.341\,\mathrm{dB}$ and LPIPS by 0.0255, with paired 95% confidence intervals excluding zero (Appendix

Table 5: Summary of reconstruction performance, computational cost, and x-free status.

Model	N_{tr}	MSE-recov (%) CIFAR-10	SVHN	FMNIST	MNIST	ImageNet	CelebA	Cost ×M1	x-free
M1	–	0						1.0	–
M3 †	1	116.87	103.86	87.09	96.2	51.54	98.73	8.1	–
M4 ($\alpha = 1.0$) †	1	52.72	68.33	61.88	63.75	63.36	63.09	2.0	✓
M5 ($\beta = 2.0$)	1	24.97	34.58	17.59	12.55	33.09	24.96	2.9	✓
M6-LO	0	-12.71	18.62	-40.63	-84.45	3.28	-30.06	4.2	–
M6	1	37.84	56.56	32.34	17.6	45.12	41.85	6.1	–
M6	2	42.42	61.05	39.20	34.32	47.72	47.21	6.3	–
M6	4	42.49	60.68	39.10	35.41	43.47	48.47	9.0	–
M6-Lobj	2	35.23	51.44	-289.58	-19294.8	44.56	-93.41	5.8	–
M6-xFree	**2**	**35.99**	**52.85**	**35.69**	**31.68**	**47.24**	**40.46**	**6.3**	✓
M7	2	43.19	63.31	45.06	44.08	30.22	45.12	12.1	–
M8	**2**	**37.65**	**53.11**	**36.28**	**32.79**	**47.03**	**40.16**	**6.9**	✓
D-optimized Reference $D(z^*)$		100						1317	–

† M3 and M4 improve reconstruction quality without enforcing decoder compatibility; their gains diminish after re-encoding through the AE (see Table 3). Bold rows highlight the learned x-free variants.

Table 6: RAE results for trajectory informativeness and pixel–perceptual reconstruction quality.

Model	N_{tr}	PSNR ↑	LPIPS ↓	rFID-5k ↓
M1	–	18.87	**0.156**	**3.91**
M4	1	17.70	0.201	5.72
M6-LO	0	18.65	0.470	25.52
M6	1	**18.99**	0.445	21.76
M6-xFree	1	18.20	0.208	5.72

D.3). Thus, despite the substantially different latent geometry, SRD again provides useful information beyond z_1 alone.

The comparison with M1 reveals a separate issue. M6 slightly improves pixel fidelity over M1 (PSNR: 18.87→18.99 dB), but substantially degrades perceptual metrics (LPIPS: 0.156→0.445; rFID-5k: 3.91→21.76). Importantly, the mismatch also occurs without trajectory conditioning; adding SRD improves both perceptual metrics over M6-LO. M6-xFree produces a more conservative refinement, improving PSNR over its pseudo-target baseline M4 (17.70→18.20 dB) while remaining much closer to M1 in perceptual quality. These results show that SRD informativeness persists in the pretrained RAE, while pixel-level reconstruction improvement and perceptual preservation are distinct objectives.

6 Discussion

6.1 SRD as a Sample-Specific Reconstruction Signal

Our central finding is that repeated self-reconstruction itself does not improve reconstruction; nevertheless, the resulting SRD contains information useful for correcting the initial reconstruction. M6 substantially outperforms the trajectory-free M6-LO, whereas assigning another sample's trajectory causes severe degradation. Thus, SRD is useful not because it captures a generic drift of the AE, but because its dynamics are tied to the individual sample.

In the convolutional AEs, successive SRD steps are locally coherent but progressively deviate from the initial direction, while remaining partially aligned with the empirical reconstruction-improving direction. This behavior is consistent with the effectiveness of simple reverse extrapolation at early steps, its limited benefit under repeated application, and the larger gains obtained by nonlinear SRD-RR. It also agrees with the observation that most useful trajectory information is captured by the first few transitions.

6.2 From Image-Space Evidence to x-Free Refinement

The large image-space gains of M3 (+6.81 dB) and M4 (+3.13 dB) show that image-space SRD contains substantial reconstruction-relevant information. However, these gains are largely lost after

re-encoding because the existing encoder does not reliably preserve the corrected reconstruction. This motivates predicting a latent correction directly and evaluating its output through the fixed decoder. M6 performs such decoder-compatible refinement, achieving a 2.01 dB average PSNR gain, but uses the original image x as its refinement target. M6-xFree instead uses the SRD-derived pseudo-target $\tilde{x}_{\mathrm{lin}}$. Once the initial self-reconstruction quantities have been generated, its refinement training and inference require no direct access to the original image x. M6-xFree achieves a 1.74 dB average PSNR gain and 40.7% average MSE recovery across six datasets, without per-sample latent optimization. These results show that substantial reconstruction improvement can be obtained using only information generated by the frozen AE's own self-reconstruction process. The auxiliary-guide variants M7 and M8 provide only marginal or dataset-dependent additional gains over M6 and M6-xFree, respectively, suggesting that the short SRD already captures much of the useful information for refinement. Detailed comparisons are provided in Appendix A.3.

6.3 Refinement Objectives and the Pretrained RAE

The RAE extension separates two questions that largely coincide in the conventional AEs: whether SRD is informative and whether the chosen refinement objective produces a desirable reconstruction. The matched M6/M6-LO comparison answers the first: trajectory conditioning improves reconstruction prediction despite the qualitatively different SRD geometry. The comparison with M1 exposes the second: optimizing pixel-space MSE can improve pixel fidelity while substantially degrading perceptual quality. Thus, SRD can remain a useful predictive signal even when the objective used to exploit it is not aligned with perceptual reconstruction quality.

A related issue appears in the conventional AE experiments. Directly regressing toward the empirically optimized latent z^* (M6-Lobj) is substantially less stable than optimizing the decoded reconstruction as in M6. This suggests that Euclidean proximity in latent space is not necessarily aligned with reconstruction quality, since decoder sensitivity can vary across latent directions. This interpretation is consistent with work on decoder-induced latent geometry, in which Euclidean latent distances need not reflect distances between decoded outputs (Arvanitidis et al., 2017). Thus, z^* serves as a useful evaluation reference, but is not necessarily an effective Euclidean regression target. More generally, reconstruction fidelity alone may not determine whether a refined latent representation remains suitable for subsequent processing. For example, StyleGAN inversion exhibits a distortion–editability trade-off, where improved reconstruction does not necessarily yield a latent representation suitable for semantic manipulation (Tov et al., 2021). Our RAE results similarly indicate that, for pretrained representation models, decoder compatibility alone may be insufficient: the refinement objective should also reflect the perceptual or downstream properties that the latent representation is expected to preserve.

6.4 Limitations and Future Work

Our main evaluation uses convolutional deterministic AEs across six datasets, while the pretrained-model extension is limited to one DINOv2-based RAE on ImageNet100. Further evaluation across broader AE families is therefore needed. The relation between SRD and decoder geometry also remains incompletely understood. Future work includes analyzing SRD through local decoder geometry, developing objectives that better balance pixel fidelity and perceptual quality, improving x-free refinement, and examining compatibility with downstream latent-space operations. More broadly, we aim to understand whether naturally occurring transient dynamics in pretrained models can serve as intrinsic, sample-specific signals for representation refinement.

7 Conclusion

We investigated whether a trained AE contains exploitable information for improving its own initial reconstruction. Self-reconstruction can degrade image fidelity while exposing dynamics useful for refinement. SRD-RR exploits these dynamics through a learned latent correction with a frozen AE and no per-sample test-time optimization. Across six datasets with convolutional AEs, nonlinear SRD-RR outperforms linear latent extrapolation, including in an x-free variant trained on an SRD-derived pseudo-target. Trajectory ablations support the practical value of correctly paired SRD features. The RAE extension further supports SRD informativeness under different latent dynamics, while showing that pixel fidelity and perceptual quality are distinct. Together, these findings establish self-reconstruction trajectories as useful signals for refinement with frozen autoencoders.

Reproducibility Statement

To support reproducibility, we report the dataset splits and evaluation protocol in Section 4.1 and Table 2. Appendix C provides detailed preprocessing, model architectures, training procedures, hyperparameters, latent normalization, and the empirical decoder-optimized reference. Appendix D provides the pretrained RAE configuration and its evaluation protocol, including perceptual metrics and confidence-interval computation.

Use of Generative AI

Generative AI tools were used to assist with language editing, literature survey, manuscript organization, code development and debugging, and discussion of experimental analyses. All AI-assisted code and analyses were reviewed and verified by the authors.

The authors made all final decisions regarding the research methodology, experimental design, interpretation of results, and conclusions, and take full responsibility for the content of this work.

References

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based StyleGAN encoder via iterative refinement. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6691–6700. IEEE, 2021.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36:41259–41282, 2023.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. *Advances in neural information processing systems*, 26, 2013.

Joaquim Campos, Simon Meierhans, Abdelaziz Djelouah, and Christopher Schroers. Content adaptive optimization for neural image compression. *arXiv preprint arXiv:1906.01223*, 2019.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *International conference on machine learning*, pp. 1078–1086. PMLR, 2018.

Antonia Creswell, Kai Arulkumaran, and Anil Anthony Bharath. Improving sampling from generative autoencoders with markov chains. *arXiv preprint arXiv:1610.09296*, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Seongmin Hong, Suh Y Jeon, Kyeonghyun Lee, Ernest K Ryu, and Young Chun. Gradient-free decoder inversion in latent diffusion models. *Advances in Neural Information Processing Systems*, 37:82982–83007, 2024.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Yoon Kim, Sam Wiseman, Andrew Miller, David Sontag, and Alexander Rush. Semi-amortized variational autoencoders. In *International Conference on Machine Learning*, pp. 2678–2687. PMLR, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jaejun Lee, Yonggyun Yu, and Hogeon Seo. Restoration of multi-channel signal loss using autoencoder with recursive input strategy. *Scientific Reports*, 15(1):13729, 2025.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, pp. 3403–3412. PMLR, 2018.

Yoshihiro Nagano, Ryo Karakida, and Masato Okada. Collective dynamics of repeated inference in variational autoencoder rapidly find cluster structure. *Scientific Reports*, 10(1):16001, 2020.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814, 2010.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Andrea Polesel, Giovanni Ramponi, and V John Mathews. Image enhancement via adaptive unsharp masking. *IEEE Transactions on image processing*, 9(3):505–510, 2000.

Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Overparameterized neural networks implement associative memory. *Proceedings of the National Academy of Sciences*, 117 (44):27162–27170, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjoern Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on graphics (TOG)*, 40(4):1–14, 2021.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN inversion: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 45(3): 3121–3138, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yibo Yang, Robert Bamler, and Stephan Mandt. Improving inference for neural image compression. *Advances in Neural Information Processing Systems*, 33:573–584, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Long Zhao, Sanghyun Woo, Ziyu Wan, Yandong Li, Han Zhang, Boqing Gong, Hartwig Adam, Xuhui Jia, and Ting Liu. Epsilon-VAE: Denoising as visual decoding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77740–77759. PMLR, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pp. 597–613. Springer, 2016.

A Additional Results and Discussion

Table A1 summarizes the main reconstruction results in PSNR and the corresponding improvements over the baseline.

A.1 Linear Image-Space and Latent-Space Correctors

Figure A1 shows validation-set PSNR sweeps for the image-space (M4) and latent-space (M5) linear extrapolation methods. For M4, we fix $\alpha = 1.0$ based on the direct linear extrapolation formulation; the validation sweep confirms that this choice is near-optimal across datasets. For M5, the validation sweeps consistently favor $\beta \approx 2.0$ across all six datasets, and we therefore use $\beta = 2.0$ for final test evaluation. Notably, the first coefficient independently learned by M5-T is also close to 2.0 across datasets.

Because M5 is training-free, computationally efficient, decoder-compatible, and x-free, we additionally examine repeated application of M5.

Figure A2 shows the performance of Iterative M5 and Iterative M5-AE, in which an AE step is applied after Iterative M5. Repeated M5 correction progressively degrades reconstruction quality. This behavior is consistent with the changing direction of the latent SRD and the limited validity of repeated linear extrapolation, motivating nonlinear SRD-RR methods such as M6 and M6-xFree.

A.2 Latent SRD Trajectories

Figure A3 visualizes the latent SRD trajectories in two dimensions using the observed magnitudes and cosine similarities of the dz_k vectors. Vector directions are fitted by least squares to reproduce the observed cosine similarities while preserving their magnitudes. The fitting error is small across all datasets, with a mean cosine-similarity error of (0.0162 ± 0.0164).

Thus, the two-dimensional representations faithfully visualize the directional changes observed in the original high-dimensional trajectories. These results are consistent with the limitations of global linear extrapolation and the degradation observed under repeated M5 correction.

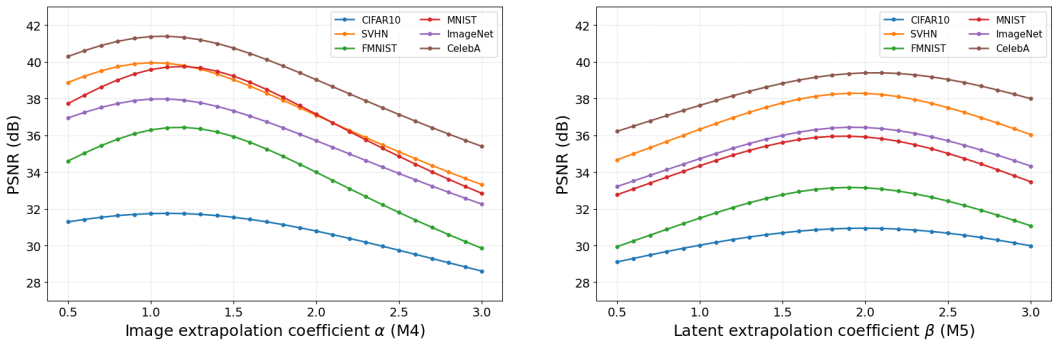
A.3 Latent-Objective and Auxiliary-Guide Ablations

Directly regressing toward the empirically optimized latent z^* (M6-Lobj) is substantially less stable than optimizing the decoded reconstruction as in M6. This suggests that Euclidean proximity in latent space is not necessarily aligned with reconstruction quality, since decoder sensitivity can vary across latent directions. Thus, z^* is useful as an evaluation reference but is not necessarily an effective Euclidean regression target.

Table A1: PSNR improvements over the baseline reconstruction and corresponding MSE recovery ratios. M1 reports the absolute PSNR, whereas all other rows report the PSNR improvement (ΔPSNR) relative to M1.

Method	N_{tr}	CIFAR-10	SVHN	FMNIST	MNIST	ImageNet	CelebA	macro ΔPSNR	macro MSE-recov. (%)	x-free
M1	–	30.42	36.91	32.37	35.33	35.03	38.33	0.00	0.00	–
M3 †	1	3.80	7.94	8.30	12.75	2.00	6.09	6.81	92.38	
M4 †	1	1.36	3.59	3.98	4.28	2.67	2.90	3.13	62.19	✓
M5	1	0.57	1.56	0.84	0.61	1.19	0.96	0.95	24.62	✓
M6-LO	0	-0.24	0.94	-1.30	-2.57	0.28	-0.77	-0.61	-24.32	
M6	1	0.95	2.80	1.74	0.86	1.66	1.60	1.60	38.55	
M6	2	1.05	3.08	2.18	1.88	1.87	2.00	2.01	45.32	
M6	4	1.06	3.06	2.19	1.96	1.93	2.07	2.04	44.93	
M6-Lobj	2	0.85	2.44	-5.46	-22.45	1.71	-2.18	-4.18	-3257.76	
M6-xFree	**2**	**0.87**	**2.53**	**1.89**	**1.66**	**1.84**	**1.66**	**1.74**	**40.65**	✓
M7	2	1.08	3.25	2.73	2.58	1.08	1.88	2.10	45.16	
M8	2	0.91	2.54	1.93	1.73	1.83	1.64	1.76	41.17	✓
$D(z^*)$	–	3.01	7.11	13.17	16.58	5.22	6.09	8.53	100.00	–

†: M3 and M4 improve image quality without enforcing decoder compatibility.



(a) α sweep (b) β sweep

Figure A1: Validation-set performance of image-space and latent-space linear extrapolation under sweeps of α and β.

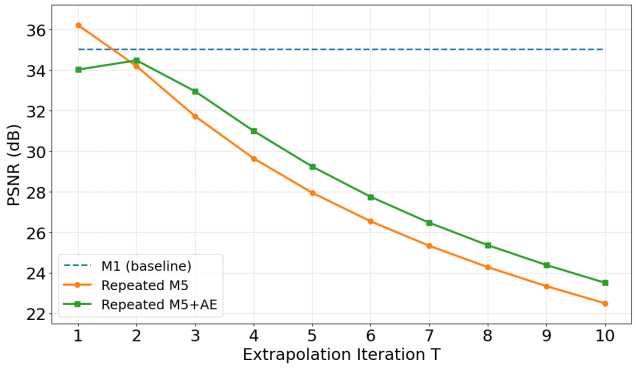


Figure A2: Performance of Repeated M5 and Repeated M5+AE on ImageNet.

The auxiliary-guide experiments (M7 and M8) examine whether additional latent guides provide useful information beyond the SRD trajectory. Under original-image supervision, M7 augments M6 with a latent guide derived from the M3 correction. Its average PSNR gain is 2.10 dB, compared with 2.01 dB for M6, while its average MSE-recov is 45.2%, essentially comparable to M6. Its effect is also dataset-dependent, including degradation on ImageNet. Thus, the additional image-derived guide is not uniformly beneficial.

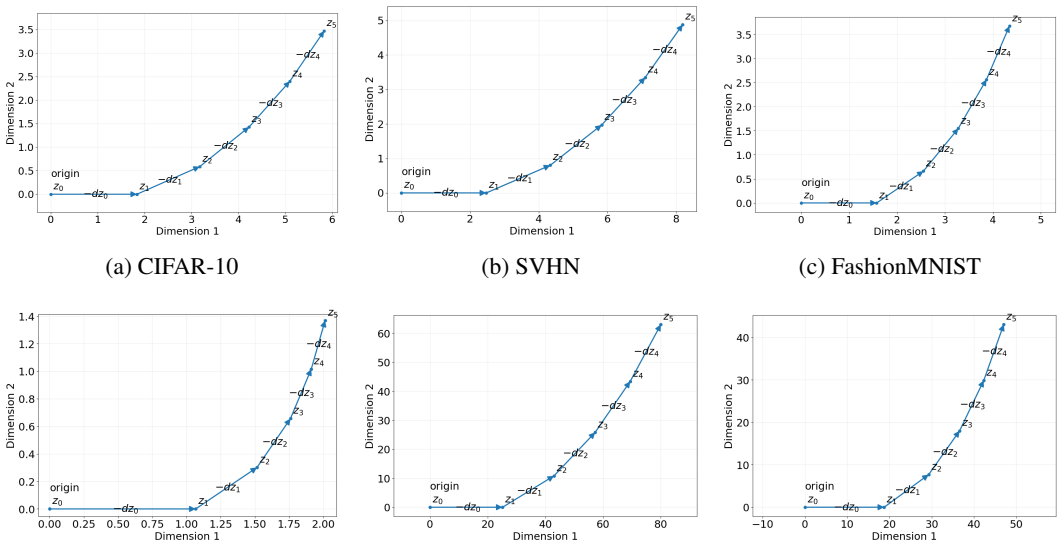
In the x-free setting, M8 augments M6-xFree with $\tilde{z}_{0,\mathrm{lin}} = E(\tilde{x}_{\mathrm{lin}})$ while using the same pseudo-target. M8 achieves an average PSNR gain of 1.76 dB and an average MSE-recov of 41.2%, compared with 1.74 dB and 40.7% for M6-xFree. The improvement is therefore small despite increasing the computational cost from $6.3\times$ to $6.9\times$ M1. This matched comparison suggests that the additional guide provides only limited benefit beyond the SRD trajectory and its SRD-derived training target.

A.4 Robustness

Figure A4 evaluates SRD-RR across five random seeds and multiple AE architectures on CIFAR-10 and ImageNet. In (a), bars show means and error bars show standard deviations across five seeds. For each seed, both the AE and SRD-RR predictor are retrained from scratch using a seed-specific training/validation split, while the test set is kept fixed. SRD-RR consistently improves reconstruction across all settings. Although the exact gains vary, the overall trend is preserved across random initializations and changes in AE depth and channel capacity.

B Qualitative Results

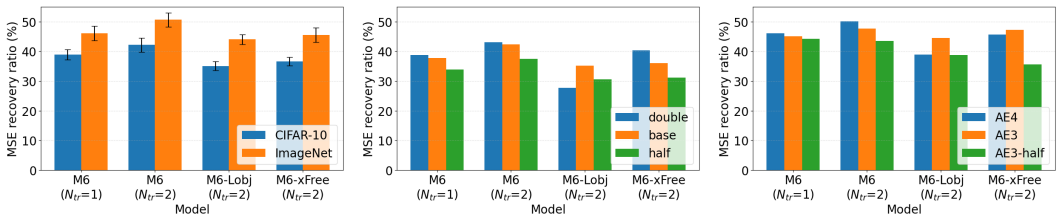
Figs. A5 and A6 show reconstruction examples and the corresponding error maps, where the errors are magnified by a factor of 50, clipped to $[0, 1]$, and colorized for visualization.



(a) CIFAR-10 (b) SVHN (c) FashionMNIST

(d) MNIST (e) ImageNet (f) CelebA

Figure A3: Changes in the latent trajectories across the six datasets.



(a) Five different seeds (b) CIFAR-10 (c) ImageNet

Figure A4: Performance consistency across (a) five random seeds and (b)(c) model architectures.

C Implementation Details for the Main Experiments

C.1 Image Pre-processing

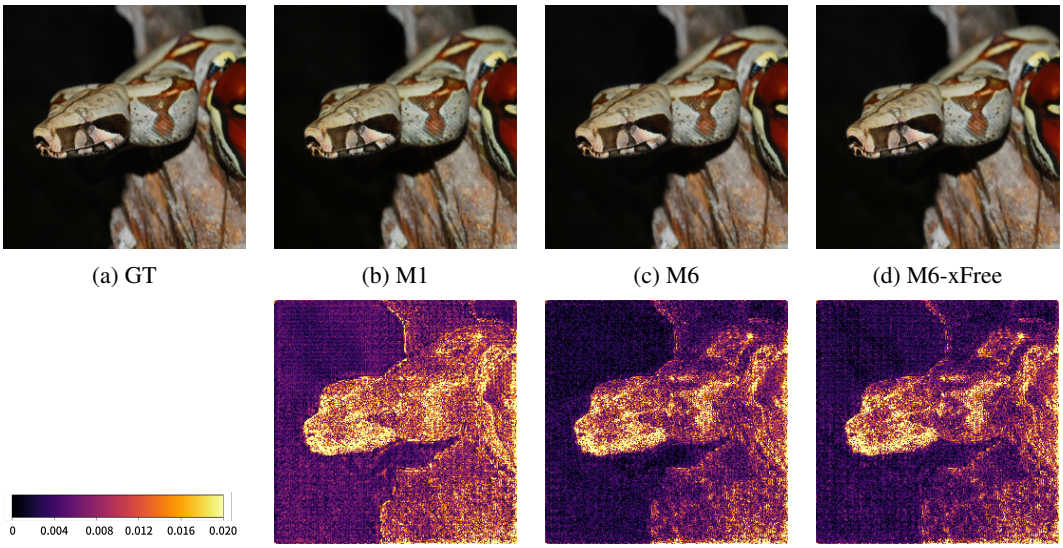
CIFAR-10 and SVHN images were used at their native 32×32 resolution, while MNIST and FashionMNIST were used at 28×28. ImageNet100 images were resized to 256×256 using bilinear interpolation. CelebA images were used at their native 218×178 resolution. All images were converted to tensors with pixel values scaled to $[0, 1]$; no additional normalization or data augmentation was applied.

Because the CelebA image dimensions are not divisible by the overall downsampling factor of the encoder, each 218×178 image was zero-padded on the bottom and right to 224×184 before being processed by the AE. After decoding, the reconstructed image was cropped back to the original 218×178 resolution before evaluation.

C.2 Autoencoder

We use fully convolutional AEs with spatial latent representations. Each encoder stage consists of a 3×3 convolution with stride 2 and padding 1, followed by batch normalization (Ioffe & Szegedy, 2015) and ReLU (Nair & Hinton, 2010). The decoder is symmetric, using transposed convolutions in reverse channel order, with a sigmoid output layer. The encoder channel dimensions are $[16, 32, 64]$ for CIFAR-10, SVHN, FMNIST, and MNIST, and $[32, 64, 128]$ for ImageNet and CelebA. The AE architecture is independent of the SRD length K.

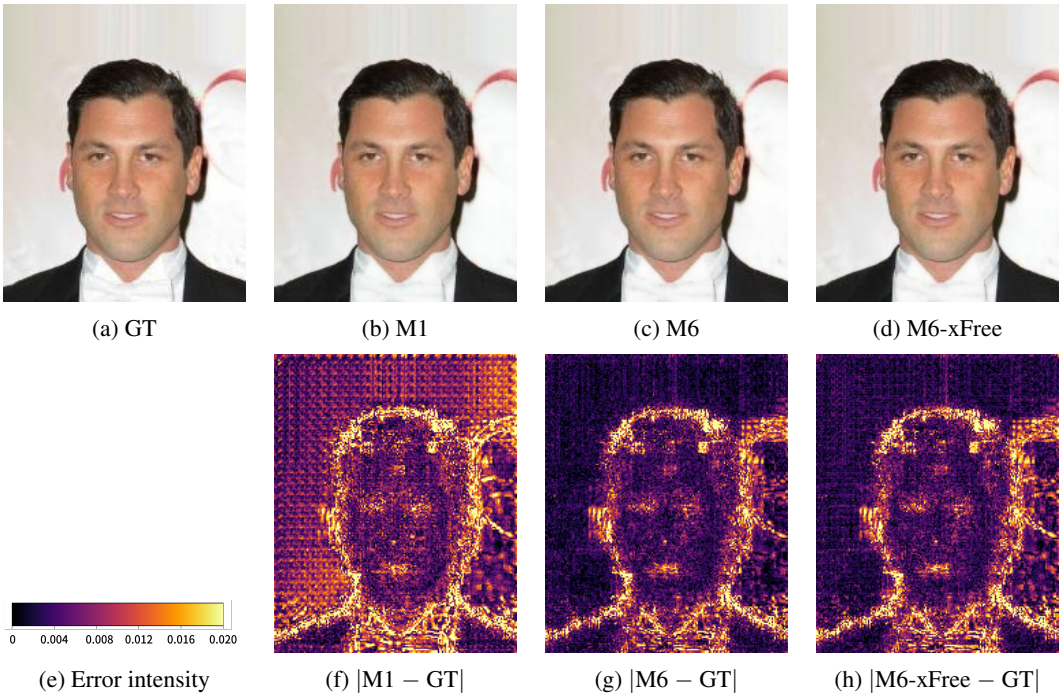
For the architecture-consistency experiments, we additionally vary the AE capacity while keeping the remaining protocol unchanged. For CIFAR-10, we use encoder channels $[8, 16, 32]$ (half), $[16, 32, 64]$ (standard), and $[32, 64, 128]$ (double). For ImageNet, we use $[16, 32, 64]$ (half),



(a) GT (b) M1 (c) M6 (d) M6-xFree

(e) Error intensity (f) $|\text{M1} - \text{GT}|$ (g) $|\text{M6} - \text{GT}|$ (h) $|\text{M6-xFree} - \text{GT}|$

Figure A5: Qualitative comparison of reconstruction results and corresponding error maps from ImageNet. M1 (37.76dB), M6 (39.71dB, MSE-recov=51.3%), and M6-xFree (39.68dB, MSE-recov=50.5%).



(a) GT (b) M1 (c) M6 (d) M6-xFree

(e) Error intensity (f) $|\text{M1} - \text{GT}|$ (g) $|\text{M6} - \text{GT}|$ (h) $|\text{M6-xFree} - \text{GT}|$

Figure A6: Qualitative comparison of reconstruction results and corresponding error maps from CelebA. M1 (38.39dB), M6 (40.97dB, MSE-recov=55.6%), and M6-xFree (40.50dB, MSE-recov=47.7%).

$[32, 64, 128]$ (standard), and a deeper four-stage architecture $[32, 64, 128, 256]$ (AE-4). Each decoder is symmetric to its corresponding encoder.

The AE is trained to minimize pixel-space MSE using Adam with a learning rate of 10^{-3} and no weight decay. We train for at most 50 epochs with early-stopping patience 5 and gradient clipping at norm 10. The batch size is 64 for the four smaller datasets and 256 for ImageNet and CelebA. Training uses FP16 automatic mixed precision (AMP) with FP32 master weights. After training, all

AE parameters and BatchNorm running statistics are kept fixed throughout SRD generation and all refinement experiments.

C.3 Image-Space Corrector

M3 concatenates the one-shot reconstruction and the observed image-space SRD,

$$[y_1, dy_1, \ldots, dy_{K-1}],$$

along the channel dimension. It consists of an initial 3×3 convolution with 128 hidden channels, three residual blocks (He et al., 2016), and a zero-initialized 3×3 output convolution. Each residual block contains two convolutions with batch normalization and a residual connection. The network predicts a residual Δx, giving

$$\tilde{x} = y_1 + \Delta x.$$

M3 is trained against the residual target $x - y_1$ using MSE. We use Adam (Kingma & Ba, 2014) with learning rate 10^{-3}, no weight decay, at most 50 epochs, early-stopping patience 6, and gradient clipping at norm 1. The batch size is 64 for the smaller datasets and 256 for ImageNet and CelebA. Training uses FP16 AMP.

C.4 Trained Linear Trajectory Corrector

M5-T learns the coefficients $\{\beta_k^*\}$ by minimizing the pixel-space MSE between $D(z_1 + \sum_k \beta_k^* dz_k)$ and the ground-truth image. We train M5-T using Adam with a learning rate of 0.03 and a batch size of 64 for at most 20 epochs, with early-stopping patience 5 and a minimum improvement tolerance of 10^{-7}.

C.5 Latent-Space SRD-RR Predictor and Hyperparameters

The nonlinear SRD-RR predictor takes the channel-wise concatenation

$$[z_1, dz_1, \ldots, dz_{K-1}]$$

as input and predicts a latent correction Δz. For CIFAR-10, SVHN, FMNIST, and MNIST, the predictor uses 64 hidden channels and four residual blocks; for ImageNet and CelebA, it uses 128 hidden channels and five residual blocks. The input and output projections use 3×3 convolutions, and each residual block consists of two 3×3 convolutions with batch normalization. The final correction layer is initialized to zero.

The corrected latent representation and reconstruction are

$$\hat{z} = z_1 + \gamma \Delta z, \qquad \hat{x} = D(\hat{z}),$$

where γ is a learnable scalar initialized to 0.5. All nonlinear SRD-RR variants use the same predictor backbone and training hyperparameters unless otherwise stated. M6-LO uses the same predictor architecture and training settings as M6, except that its input projection receives only z_1 and therefore has fewer input channels.

For supervised SRD-RR, the predictor minimizes pixel-space MSE between $D(\hat{z})$ and the original image x. The x-free variant instead uses the SRD-derived pseudo-target

$$\tilde{x}_{\text{lin}} = y_1 + (y_1 - y_2)$$

and minimizes the same pixel-space MSE with $\tilde{x}_{\text{lin}}$ as the target. Once the self-reconstruction quantities have been generated, M6-xFree requires no direct access to the original image x during refinement training or inference. For the x-free variants (M6-xFree and M8), early stopping uses only the pseudo-target validation loss. We optimize the SRD-RR predictors using AdamW (Loshchilov & Hutter, 2017) with learning rate 5×10^{-4} and weight decay 10^{-3}. Training is performed for at most 150 epochs with early-stopping patience 10 and gradient clipping at norm 1. The batch size is 64 for the smaller datasets and 256 for ImageNet and CelebA. Pixel-objective predictors are trained using FP16 AMP, whereas the latent-objective M6-Lobj is optimized in FP32 for numerical stability.

Detailed hyperparameters are summarized in Table A2.

Table A2: Training hyperparameters for the learned components. Batch sizes are shown as small/large datasets.

	AE	M3	SRD-RR
Optimizer	Adam	Adam	AdamW
Learning rate	10^{-3}	10^{-3}	5×10^{-4}
Weight decay	0	0	10^{-3}
Batch size	64 / 256	64 / 256	64 / 256
Max. epochs	50	50	150
Early-stop patience	5	6	10
Gradient clipping	10	1	1
Objective	pixel MSE	residual MSE	pixel MSE
Precision	FP16 AMP	FP16 AMP	FP16 AMP†

M6-Lobj uses a latent-space MSE objective and FP32 optimization.

C.6 Latent Normalization

The latent representation z_1 and trajectory differences dz_k are standardized per dimension using statistics estimated from the training set. Statistics for z_1 are estimated separately, while the dz_k share common per-dimension normalization statistics across trajectory transitions. The predicted correction is rescaled using the trajectory-difference statistics before being added to z_1. The same normalization is used during training and inference.

For trajectory storage, z_1 is stored in FP16, whereas dz_k is stored in FP32. Other latent states are reconstructed from these quantities when required.

C.7 Empirical Decoder-Optimized Reference

For each sample, the empirical decoder reference is obtained by solving

$$z^* = \arg\min_z \|D(z) - x\|_2^2$$

with the decoder fixed and initialization $z_0 = E(x)$. We use Adam with learning rate 0.01 for at most 500 optimization steps, gradient clipping at 14, and early stopping. No latent-anchor regularization is used. Optimization is performed in FP32 and batched only for computational efficiency.

In all evaluated samples, the optimized reference $D(z^*)$ achieved lower reconstruction error than the one-shot baseline y_1; no sample with a non-positive MSE-recov denominator was observed. This optimization is used to construct the empirical evaluation reference and, for M6-Lobj, the latent regression target z^* on the training data. For test samples, z^* is used only for evaluation and analysis. No per-sample latent optimization is required by SRD-RR at inference time.

D Extension to a Pretrained Representation Autoencoder

D.1 Experimental Setup

To examine whether SRD extends beyond the convolutional deterministic AEs used in the main experiments, we additionally evaluate a pretrained DINOv2-based Representation Autoencoder (RAE) (Zheng et al., 2026). We use the publicly available `RAE-dinov2-wReg-base-ViTXL-n08` checkpoint and keep both the encoder and decoder frozen. Input images are bicubically resized and center-cropped to 256×256, producing a $768 \times 16 \times 16$ latent representation.

The main experimental settings are summarized in Table A3.

We use fixed ImageNet100 subsets of 20,000 training, 2,000 validation, and 5,000 test images. The nonlinear corrector follows the architecture in Appendix C, with 128 hidden channels and five residual blocks (as in the main ImageNet experiments).

We compare M1, M2 (y_2, y_3, y_5), M4, M5, M6-LO, M6 ($N_{\mathrm{tr}} = 1, 2$), and M6-xFree ($N_{\mathrm{tr}} = 1, 2$). The M4 coefficient is fixed to $\alpha = 1$ as in the main experiments, whereas validation selected $\beta = 1$ for M5 before final test evaluation.

Table A3: Experimental settings for the pretrained RAE extension.

Setting	Value
RAE checkpoint	DINOv2-based RAE, ViT-XL decoder
Input resolution	256×256
Latent shape	$768 \times 16 \times 16$
RAE parameters	Frozen
Dataset	ImageNet
Predictor train / val / test	20,000 / 2,000 / 5,000
Predictor hidden channels	128
Residual blocks	5
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	10^{-3}
Batch size	64
Max. epochs / patience	30 / 8
Gradient clipping	1.0
γ initialization	0.5
M4 coefficient α	1.0
M5 coefficient β	1.0
Evaluated N_{tr}	1, 2
Metrics	MSE, PSNR, SSIM, LPIPS, rFID-5k

Table A4: Latent SRD geometry of the pretrained RAE. Values are mean $\pm$ standard deviation over validation images.

k	0	1	2	3	4
$\|dz_k\|_2$	197.83±27.02	184.87±29.76	192.06±34.16	200.25±36.00	203.22±33.75
$\cos(dz_k, dz_{k+1})$	0.094±0.067	0.173±0.073	0.173±0.081	0.163±0.088	–
$\cos(dz_0, dz_k)$	–	0.094±0.067	−0.084±0.043	−0.089±0.034	−0.072±0.028

We report MSE, PSNR, SSIM, LPIPS, and rFID-5k. The first four metrics are computed per sample. LPIPS is computed using the pretrained AlexNet-based LPIPS model (Zhang et al., 2018), with images normalized to $[-1, 1]$, and averaged over the 5,000 test images. rFID-5k is computed between the same fixed 5,000 ground-truth and reconstructed images using CleanFID in `legacy_tensorflow` mode. These rFID values are intended for comparisons within this experiment and should not be directly compared with rFID values computed using a different number of images. For M6 versus M6-LO, 95% confidence intervals for mean metric differences are computed using paired t intervals over the same 5,000 test images.

D.2 Distinct Latent Geometry of SRD in the pretrained RAE

To examine whether the latent SRD observed in the convolutional AEs persists in a substantially different representation space, we analyze repeated self-reconstruction trajectories of the pretrained RAE. Using validation images, we measure the norm of each reverse-step vector dz_k, the cosine similarity between successive steps, and the cosine similarity between each step and the initial direction dz_0. As in Sec. 4.4, dz_0 is used only for geometric analysis and is not available to the x-free refinement models.

Table A4 summarizes the latent SRD statistics. To visualize the corresponding trajectory geometry, we fit a three-dimensional trajectory to the measured step magnitudes and cosine similarities, following the same procedure as in the main experiments, as illustrated in Fig. A7. Although the two-dimensional fitting resulted in relatively large errors, the three-dimensional fit reproduces all measured cosine similarities with an average error below 10^{-10}. Thus, the figure provides an accurate three-dimensional representation of the measured trajectory geometry.

The RAE exhibits a markedly different SRD geometry from the convolutional AEs in the main experiments.

First, the step magnitude does not progressively decay: after decreasing from $\|dz_0\|_2 = 197.8$ to $\|dz_1\|_2 = 184.9$, it increases again and reaches 203.2 at dz_4, showing a markedly different magnitude profile from the convolutional AEs.

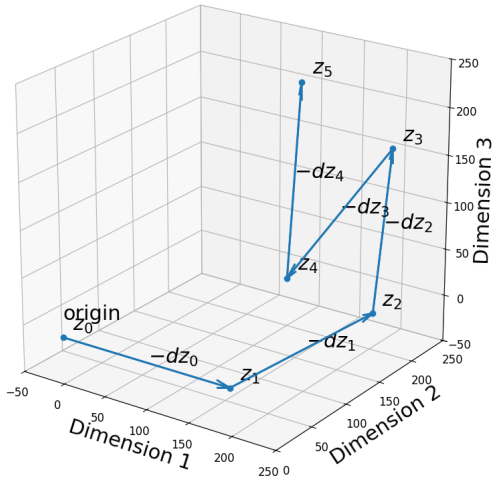


Figure A7: 3D visualization of latent SRD trajectories in the pretrained RAE on ImageNet100.

Table A5: Reconstruction performance of SRD-based refinement on the pretrained RAE. MSE, PSNR, SSIM, and LPIPS are averaged over the fixed 5,000 test images; rFID-5k is computed over the same complete test set. Best values are shown in bold.

Model	N_{tr}	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	rFID-5k ↓
M1	–	0.01602	18.87	0.4696	**0.1558**	**3.91**
M2 (y_2)	1	0.02124	17.41	0.4286	0.2193	6.29
M2 (y_3)	2	0.02612	16.37	0.4002	0.2770	11.21
M2 (y_5)	4	0.03516	14.97	0.3601	0.3740	33.44
M4	1	0.02153	17.70	0.4245	0.2009	5.72
M5	1	0.01911	17.94	0.4363	0.1943	4.74
M6-LO	0	0.01584	18.65	0.4899	0.4701	25.52
M6	1	**0.01483**	**18.99**	**0.4948**	0.4446	21.76
M6	2	0.01494	18.94	0.4927	0.4527	22.66
M6-xFree	1	0.01791	18.20	0.4532	0.2083	5.72
M6-xFree	2	0.01800	18.17	0.4524	0.2108	6.00

Second, successive directions exhibit only weak positive alignment. Their mean cosine similarity is 0.094 for (dz_0, dz_1) and approximately 0.16–0.17 thereafter, in contrast to the strong local directional coherence observed for the convolutional AEs in Fig. 1a.

Third, the trajectory rapidly departs from its initial direction. While dz_1 remains weakly aligned with dz_0, the mean cosine similarity with dz_0 becomes negative from dz_2 onward, reaching -0.084, -0.089, and -0.072 for dz_2, dz_3, and dz_4, respectively.

Taken together, SRD persists in the pretrained RAE but has a qualitatively different geometry: instead of the locally coherent and gradually curving trajectories observed in the convolutional AEs, the RAE exhibits weak local directional alignment, rapid departure from the initial direction, and non-decaying step magnitudes. This strongly curved geometry also provides a natural explanation for why fixed linear latent extrapolation is less effective in the RAE, as examined next.

D.3 SRD-Guided Reconstruction Refinement in the Pretrained RAE

Table A5 summarizes the reconstruction results on the pretrained RAE. Repeated self-reconstruction again progressively degrades fidelity: from M1 to y_5, PSNR decreases from 18.87 to 14.97 dB, while rFID-5k increases from 3.91 to 33.44. Thus, the degradation under repeated application observed for the convolutional AEs also occurs in the pretrained RAE, despite its substantially different encoder–decoder architecture and latent representation. Because PSNR is computed independently for each sample before averaging, its ordering need not exactly match that of the dataset-averaged MSE.

The linear SRD baselines M4 and M5 do not improve over M1, achieving 17.70 and 17.94 dB, respectively, compared with 18.87 dB for M1. The limited effectiveness of fixed linear latent ex-

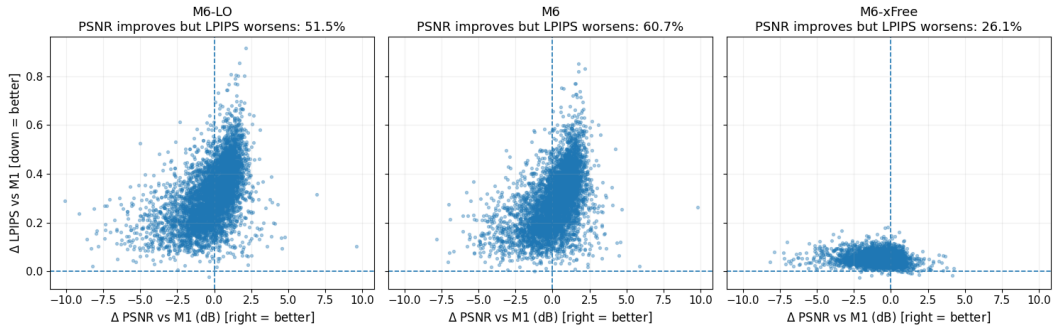


Figure A8: Per-sample pixel–perceptual discrepancy under RAE refinement.
Positive ΔPSNR indicates improved pixel fidelity, whereas positive ΔLPIPS indicates degraded perceptual similarity. The frequent occurrence of samples with improved PSNR but worse LPIPS shows that pixel-level improvement does not necessarily imply perceptual improvement. M6 and M6-xFree use $N_{\mathrm{tr}} = 1$.

trapolation is consistent with the weak local alignment and rapid directional changes of the RAE trajectory observed in Appendix D.2.

In contrast, nonlinear SRD conditioning remains informative. The trajectory-free M6-LO achieves 18.65 dB, whereas adding the first observable SRD transition improves PSNR to 18.99 dB for M6 ($N_{\mathrm{tr}} = 1$) and reduces mean MSE from 0.01584 to 0.01483. The mean paired differences (M6 - M6-LO) are $\Delta\mathrm{PSNR} = +0.341\,\mathrm{dB}$ (95% CI: $[0.310, 0.373]$) and $\Delta\mathrm{LPIPS} = -0.0255$ (95% CI: $[-0.0262, -0.0248]$). M6 also improves SSIM and rFID-5k relative to M6-LO (Table A5). Thus, despite the qualitatively different latent dynamics, explicitly providing SRD information consistently improves prediction over the matched trajectory-free baseline. A second transition provides no further gain, suggesting that the useful SRD horizon depends on the underlying encoder–decoder representation.

M6-xFree shows a complementary result. With $N_{\mathrm{tr}} = 1$, it improves PSNR from 17.70 dB for its SRD-derived linear pseudo-target baseline M4 to 18.20 dB, showing that SRD-conditioned latent correction can recover part of the loss introduced by the pseudo-target while remaining decoder-compatible. It remains below M1 in pixel fidelity, and an additional transition again provides no further benefit. The distinct pixel–perceptual behavior of M6 and M6-xFree is analyzed next.

D.4 Pixel–Perceptual Trade-off and Qualitative Analysis

The pretrained RAE reveals a clear distinction between the usefulness of SRD information and the perceptual effect of the refinement objective. As shown above, adding SRD to M6-LO improves all five evaluated metrics. However, comparison with the original one-shot reconstruction M1 reveals a different issue: optimizing the latent correction for pixel-space MSE does not necessarily preserve perceptual reconstruction quality.

For M6 ($N_{\mathrm{tr}} = 1$), mean PSNR improves from 18.87 to 18.99 dB relative to M1, MSE decreases from 0.01602 to 0.01483, and SSIM increases from 0.4696 to 0.4948. In contrast, both perceptual metrics worsen substantially: LPIPS increases from 0.1558 to 0.4446, and rFID-5k from 3.91 to 21.76. Thus, improved pixel fidelity under latent refinement does not necessarily imply improved perceptual similarity in the pretrained RAE. The mismatch also occurs without SRD conditioning: M6-LO has worse LPIPS and rFID-5k than M6 under the same pixel-MSE objective. Adding SRD therefore improves both perceptual metrics relative to the trajectory-free baseline.

M6-xFree exhibits a substantially different trade-off. Although its PSNR is lower than that of supervised M6 (18.20 vs. 18.99 dB), its LPIPS and rFID-5k are much closer to M1 (0.2083 and 5.72, respectively). It also largely preserves the perceptual characteristics of its pseudo-target M4: rFID-5k remains 5.72, while LPIPS changes from 0.2009 to 0.2083. Thus, the SRD-derived pseudo-target leads to a more conservative refinement that preserves perceptual structure substantially better than direct ground-truth pixel-MSE optimization.



(a) GT

(b) M4
PSNR=14.89dB
LPIPS=0.2188

(c) M6-LO
PSNR=16.66dB
LPIPS=0.4711



(d) M1
PSNR=16.41dB
LPIPS=0.1788

(e) M6-xFree
PSNR=16.49dB
LPIPS=0.2297

(f) M6
PSNR=17.51dB
LPIPS=0.4545

Figure A9: Qualitative comparison on ImageNet100 using the pretrained RAE.
M6 improves PSNR but worsens LPIPS relative to M1, while M6-xFree retains more fine-scale detail and achieves lower LPIPS than M6.

Figure A8 shows that improved pixel fidelity is frequently accompanied by worse perceptual similarity at the sample level. A similar pattern is observed for M6-LO, indicating that this behavior is not specific to SRD conditioning.

The qualitative examples in Figs. A9 and A10 illustrate the same behavior. M6 can reduce pixel error while smoothing fine textures and structural detail, whereas M6-xFree generally retains more of the perceptual structure present in M1. These results separate two conclusions: SRD remains useful for predicting latent corrections in the pretrained RAE, but the objective used to exploit this information determines whether improved pixel fidelity translates into perceptually desirable reconstruction.



(a) GT

(b) M4
PSNR=15.73dB
LPIPS=0.1896

(c) M6-LO
PSNR=17.38dB
LPIPS=0.4901



(d) M1
PSNR=15.92dB
LPIPS=0.1480

(e) M6-xFree
PSNR=17.35dB
LPIPS=0.1688

(f) M6
PSNR=15.28dB
LPIPS=0.4767

Figure A10: Qualitative comparison on ImageNet100 using the pretrained RAE.
M6 deteriorates in both PSNR and LPIPS relative to M1.
In contrast, M6-xFree improves PSNR over M1 with a modest increase in LPIPS and substantially better perceptual similarity than M6.